J. Child Lang., Page 1 of 30. © Cambridge University Press 2011
doi:10.1017/S0305000911000171

How to measure the onset of babbling reliably?*

INGE MOLEMANS, RENATE VAN DEN BERG, LIEVE VAN SEVEREN AND STEVEN GILLIS
University of Antwerp

(*Received* 19 *June* 2010 – *Revised* 12 *February* 2011 – *Accepted* 10 *April* 2011)

ABSTRACT

Various measures for identifying the onset of babbling have been proposed in the literature, but a formal definition of the exact procedure and a thorough validation of the sample size required for reliably establishing babbling onset is lacking. In this paper the reliability of five commonly used measures is assessed using a large longitudinal corpus of spontaneous speech from forty infants (age 0;6–2;0). In a first experiment it is shown that establishing the onset of babbling with reasonable (95%) confidence is impossible when the measures are computed only once, and when the number of vocalizations are not equal for all children at all ages. In addition, each measure requires a different minimal sample size. In the second experiment a robust procedure is proposed and formally defined that permits the identification of the onset of babbling with 95% confidence. The bootstrapping procedure involves extensive resampling and requires relatively few data.

INTRODUCTION

In every domain of science that relies on measuring phenomena in empirical data, there is a need for measuring reliably. In the domain of language acquisition, in which a primary research method consists of collecting naturalistic observational data, the use of valid methods for reliably measuring particular phenomena in those naturalistic observations should be a matter of great concern. Although phrasing this concern may sound like stating a truism, Tomasello and Stahl (2004: 101) start their article with the observation that '[T]here has been relatively little discussion in the field of child language acquisition about how to best sample from children's

[*] Preparation of this paper was supported by a grant of the Research Council of the University of Antwerp, and a grant of the Flemish Research Council FWO. The authors thank the children and their parents who generously participated in this study. Thanks are also due to two anonymous reviewers and the JCL action editor for many valuable suggestions and comments.

spontaneous speech, particularly with regard to quantitative issues'. In that paper, the authors focus on – among other things – the size of the sample and the periodicity of sampling in longitudinal studies. They cogently argue that if a phenomenon has a certain incidence 'in the real world', a particular sample size is required in order to capture that phenomenon with a particular degree of confidence in observational data. For instance, suppose segment /x/ has an incidence of 1/100 (it occurs once every 100 tokens) and segment /y/ has an incidence of 1/1,000. It is straightforward to see that in a sample of, say, 100 tokens segment /x/ is expected to occur at least once, and segment /y/ is not expected to occur at all. Now suppose – for the sake of the argument – that a researcher collects one hour of observational data from a normally developing child and videotapes a so-called late talker also for one hour. The sample of the former may consist of 1,000 tokens of segments, while the sample of the latter may consist of only 100 tokens, not surprisingly because late talkers are well known to be less voluble or talkative. The researcher analyzes both samples and observes that in both samples the frequent segment /x/ occurs, but that segment /y/, the segment with a low incidence, only appears in the speech of the normally developing child and not in that of the late talker. The (hypothetical) researcher concludes from this observation that late talkers produce less low-frequency segments, and may develop a theoretical account of why this is the case, and may even formulate the clinical implications of this observation. But in this example the researcher's observation is simply due to a difference in the size of the samples that were analyzed, and the theoretical and clinical implications that our hypothetical researcher draws from them may be completely erroneous because of a methodological flaw.

The evaluation of sample size issues in the computation of formal measures of language acquisition and development was highlighted quite recently by, among others, Tomasello and Stahl (2004) and Rowland, Fletcher and Freudentahl (2008). But scattered throughout the literature are reports that phrase a similar concern with respect to often used measures such as MLU (mean length of utterance; Klee & Fitzgerald, 1985) and type/token ratio (Richards, 1987; Malvern, Richards, Chipere & Durán, 2004). The basic message is that if a measure is applied to the data of two different children, the yardstick used for measuring should be the same in the two cases: the sample sizes should be equal as well as the unit in which those sizes are measured. For instance, lexical diversity crucially depends on the size of the sample, and that size may be expressed in terms of utterances or in terms of words, leading to very diverse results (Hutchins, Brannick, Bryant & Silliman, 2005). Only if measures are formally defined and if they can be reliably applied, can the outcomes for multiple participants in a study be compared. Moreover, these formal requirements are a condition sine qua none for comparing the results of different studies that

claim to measure the same phenomenon. In this paper we will address these questions with regard to measures that have been proposed over the years to compute the age at onset of babbling on the basis of samples of spontaneous prelexical vocalizations.

There is general agreement in the literature that for typically developing children the onset of babbling occurs before age 0;11 (e.g. Koopmans-van Beinum & van der Stelt, 1986; Nathani, Ertmer & Stark, 2006; Oller, 1980; Roug, Landberg & Lundberg, 1989; Stark, 1980). A delayed onset of babbling is even considered a predictor of later speech and language problems (Oller, Eilers, Neal & Schwartz, 1999). For instance, hearing-impaired infants start babbling considerably later than typically developing infants and they are frequently found to have a deviant speech and language development (Koopmans-van Beinum, Clement & van den Dikkenberg-Pot, 2001; Oller & Eilers, 1988; Oller, Eilers, Bull & Carney, 1985; Stoel-Gammon & Otomo, 1986). The age at which infants start babbling is used in research and in clinical practice as a very early diagnostic marker of speech and language development. Care should therefore be taken regarding the reliability of sampling procedures with which babbling onset measures are applied, and regarding the comparability of babbling onset results with different measures for the same child.

In the literature several procedures have been proposed for identifying the onset of babbling in samples of spontaneously produced prelexical vocalizations. Perhaps the most well known is the one of Oller and Eilers (1988). They view the onset of babbling as the emergence of mature-sounding or canonical syllables in the infant's vocal output, i.e. syllables with a fully resonant nucleus (a vowel) and at least one consonantal, non-glottal margin (Oller, 2000). Babbling onset is credited if the percentage of canonical syllables computed over all utterances in the entire sample equals or exceeds 20%. Oller and Eilers (1988) baptized this measure the 'Canonical Babbling Ratio' (henceforth: CBR^{utt}), and define it as in (1).

$$CBR^{utt} = \frac{\#\text{of canonical syllables}}{\text{Total \# of utterances}} \quad (1)$$

Oller and Eilers (1988: 444) report that for the identification of babbling onset for laboratory purposes a minimal sample size of 50 utterances 'is commonly sought', excluding vegetative or reflexive utterances from the sample. This sample size was deemed adequate for the calculation of CBR^{utt} by Rvachew, Creigthon, Feldman and Sauve (2001), who found high correlations between CBR^{utt} values calculated on the first 50 utterances and calculated on 100 consecutive utterances of the same speech sample. For the 21 typically developing, normally hearing subjects in Oller and Eilers' study the age at onset of canonical babbling lay between 0;6 and 0;10

(mean: 7·6 months), while nine deaf subjects failed to start babbling before 0;11 (range 0;11 – 2;1).

In the course of the 1990s the formula for the canonical babbling ratio was slightly changed (e.g. Oller, Eilers, Steffens, Lynch & Urbano, 1994). Instead of computing the ratio of syllables over utterances, the proposal was to compute the ratio of the number of canonical syllables over the total number of syllables in the sample, giving the CBR^{syl} score defined in (2).

$$CBR^{syl} = \frac{\#\text{of canonical syllables}}{\text{Total } \#\text{of syllables}} \tag{2}$$

This change of the procedure ascertained equal potential ranges for the nominator and the denominator in the equation. Values obtained for CBR^{syl} are slightly lower than those of CBR^{utt}, and therefore the critical value indicating babbling onset was lowered to 0·15 for CBR^{syl}. Lynch, Oller, Steffens, Levine, Basinger and Umbel (1995) used this CBR^{syl} criterion for evaluating babbling onset in thirteen infants with Down Syndrome as compared to twenty-seven typically developing children. Results based on samples of at least seventy vocalizations – again excluding vegetative and reflexive sounds – showed that babbling onset for typically developing children occurred between 0;6 and 1;0 (mean: 8·1 months), while the age range for Down Syndrome children was between 0;6 and 1;2 (mean: 10 months).

In the original conception of the CBR, only non-glottal consonants were considered as margins of a canonical syllable. Inspired by Stoel-Gammon (1989), Chapman, Hardin-Jones, Schulte and Halter (2001) narrowed down the definition of a canonical syllable to include only syllables containing combinations of vowels with TRUE consonants. True consonants are, according to Stoel-Gammon (1989), all consonants except glottals and glides. Apart from excluding syllables containing exclusively glottal consonants, those containing glides are also not counted in the computation of what can be called the True Canonical Babbling Ratio (TCBR). Analogous to the CBR, the TCBR can be computed with the total number of utterances ($TCBR^{utt}$, as in (3)) or the number of syllables ($TCBR^{syl}$, as in (4)) in the denominator.

$$TCBR^{utt} = \frac{\#\text{of true canonical syllables}}{\text{Total } \#\text{of utterances}} \tag{3}$$

$$TCBR^{syl} = \frac{\#\text{of true canonical syllables}}{\text{Total } \#\text{of syllables}} \tag{4}$$

A ratio of 0·20 or greater for $TCBR^{utt}$ is considered as the threshold for the canonical babbling stage. In practice, studies applying this criterion have worked with the $TCBR^{syl}$, as in (4), with the cut-off point for babbling

onset at a ratio of 0·15 or greater. Chapman *et al*. (2001), for example, studied prelexical vocal development at the age of 0;9 in 30 infants with unrepaired cleft palate and 15 age-matched non-cleft controls. They counted the number of infants in each group who had reached a $TCBR^{syl}$ of 0·15 or greater. Only 14 out of 30 cleft palate infants could be credited with babbling onset, while 14 out of 15 controls without cleft palate were babbling at that age. No sample size restrictions were reported, but the recordings used in the study lasted approximately one hour or until at least 100 utterances had been sampled (p. 1272).

The babbling onset measures introduced so far focus on the production of canonical syllables as the crucial aspect of babbled utterances. Another notable hallmark of babbling is its multisyllabic, often reduplicated nature. Van der Stelt & Koopmans-van Beinum (1986) defined babbling as the production of vocalizations with continuous or interrupted phonation in combination with multiple articulatory movements in the course of one breath unit. In their study they instructed parents to identify the onset of this type of behaviour in their infant's vocal output: 'The parents were to report the day on which they recognized babbling from their baby for the first time' (p. 164). Parents of 51 children reported the onset of babbling as occurring between the ages of 0;4 and 0;11 (mean age: 0;7). Likewise, Fagan (2009: 504) determined the onset of reduplicated babbling 'by parent report of two vocalizations containing CV syllable repetition and evidence of syllable repetition on audio or videotape'. She found a mean age of reduplicated babbling onset at 7·1 months in the 18 children in her study (range: 4·5–12 months).

A more formal criterion for identifying babbling onset in a fragment of vocal output and based on the multisyllabicity requirement was suggested by Schauwers, Gillis, Daemers, De Beukelaer and Govaerts (2004). They collected monthly samples, and took the session in which a child for the first time evidenced two or more prelexical vocalizations with interrupted or uninterrupted phonation and multiple articulatory movements as indicative of babbling onset, provided that the number of babbled utterances did not decrease to less than two in the following two months (p. 266). This criterion for determining the onset of babbling will be called the multisyllabicity criterion (henceforth: MULTI). The multisyllabicity criterion does not only introduce a different angle on the description of children's vocal productions, it also introduces a longitudinal aspect: vocalizations from three consecutive months are to be taken into account instead of those from a single month. Schauwers *et al*. used selections of 20 relatively 'voluble' minutes out of the original recording of at least an hour, without further restrictions on sample size. Babbling onset for the 10 typically developing, normally hearing children in this study was credited between the ages of 0;6 and 0;8 (median: 0;6).

In the studies reviewed, different methodologies were used and divergent criteria for establishing the onset of babbling were applied. Nonetheless, all studies agree on the finding that babbling onset typically occurs in the second half of the first year. This poses a number of interesting problems that will be dealt with in this paper. First of all, does applying the different criteria to the same data lead to identifying the onset of babbling at exactly the same point in time? For instance, Chapman *et al.* (2001) used the $TCBR^{syl}$ criterion and Schauwers *et al.* (2004) used the MULTI criterion. Both studies also report results for babbling onset based on the CBR^{syl}/CBR^{utt} criterion. However, they do not provide information about the comparability of the ages at babbling onset for individual children with each of the measures.

A second important issue, which logically precedes the previous one, concerns the size of the sample needed to reliably compute the different measures of babbling onset. Up till now that sample size has not been thoroughly validated for any of the criteria. Rvachew *et al.* (2001) found high correlations among CBR^{utt} values calculated on a sample of 50 utterances and calculated on a sample of 100 utterances. However, they do indicate that there was a mean difference between the two CBR^{utt} values of 0·05. With the threshold for babbling onset at a CBR^{utt} value of 0·20, it remains unclear which sample size suffices to establish reliably whether a child scores under or above that threshold.

Sampling issues should be made a matter of great concern in the evaluation of babbling onset in order to increase the reliability, validity and comparability of results, as in other domains of child language research (Hutchins *et al.*, 2005; Rowland *et al.*, 2008; Tomasello & Stahl, 2004). However, in the studies reviewed, sample size was not an issue at stake: either an arbitrary number of utterances or an arbitrary number of minutes of recording was taken as the yardstick. But whether this yardstick led to a reliable measurement of the onset of babbling was not assessed. Moreover, not even all studies have ascertained that sample sizes were equal for all children. A comparison of age at babbling onset between children on the basis of unequal sample sizes may not be justified: the chance of finding 2 multisyllabic utterances in the course of a session containing 500 utterances is much larger than finding them in the course of a session containing only 50 utterances. The same holds for the criteria involving the calculation of a ratio of well-formed syllables to utterances or syllables (CBR and TCBR): the size of the denominator influences the results. Two children who produce the same absolute number of canonical syllables in the course of a session can still obtain very dissimilar canonical babbling ratios if that session contained 50 utterances for one child and 500 utterances for the other. It is very important to include equal amounts of data for all children and sessions in the procedure. This avoids differences between children in

the result of a certain measure that are caused solely by differences in the volubility of those children. In this paper we will address the sample size issue for each of the five criteria for babbling onset outlined in this 'Introduction'.

The aim of this paper is to answer two fundamental questions. First of all, we investigate whether a minimal sample size can be established that permits the reliable identification of the age at onset of babbling for each of the measures ((T)CBR$^{utt/syl}$ and MULTI). This question will be explored in Experiments 1 and 2. A large longitudinal corpus of spontaneous speech from forty infants is used as an empirical database. In the first experiment it is investigated whether for each measure a minimal sample size can be established that permits identifying the onset of babbling reliably by computing the measure only once. In the second experiment, a more lenient method will be proposed that makes use of multiple randomly drawn samples and the computation of an average value. The second research question explores whether the various measures yield comparable results. Do all five measures indicate the same age for the onset of babbling? This question will be answered on the basis of the babbling onset results for each of the measures that were reached with a reliable sample size and procedure.

EXPERIMENT 1

METHOD

Participants

Data were collected from a group of forty typically developing children living in the Dutch-speaking part of Belgium. All of these children were raised in monolingual homes, acquiring the standard variant of Dutch. Typical development was established through parent case history report and the administration of a checklist of the attainment of communicative and motor milestones (largely based on the checklist developed by Kind en Gezin, the Flemish infant welfare centre). Normal language development was monitored by the administration of the Dutch version of the CDI at 1;0, 1;6 and 2;0 (*N-CDI*; Zink & Lejaegere, 2002).

Data collection and transcription

Monthly recordings were made in the children's home environments of spontaneous interactions between the children and their caretakers. The children were followed longitudinally from the age of 0;6 onwards up to 2;0. The recordings were made using a JVC digital video camera, and each recording lasted 60–90 minutes.

From each recording a selection of approximately 20 minutes was made by a member of the research team. The aim was to select uninterrupted

stretches of delineated interactions in which the child was vocally active. These selections were transcribed according to the CHAT conventions (MacWhinney, 2000).

The children's lexical utterances were orthographically and phonemically transcribed. The coding of the children's prelexical utterances was done according to the procedure used in Schauwers *et al.* (2004). Non-vegetative, non-reflexive comfort sounds uttered within one breath unit and without a consistent sound—meaning relationship were considered prelexical utterances. Each prelexical utterance was coded on a number of different dimensions (coding tiers), three of which are of particular importance for the present paper:

(a) Each utterance was characterized in terms of a combination of phonation (no phonation, uninterrupted or interrupted phonation) and number of articulatory movements (no articulation or a sustained vowel, one articulation, or two or more articulations), in accordance with the sensorimotor model developed by Koopmans-van Beinum and van der Stelt (1986). This coding allows the automatic distinction of babbles from other types of prelexical vocalizations, as required for MULTI.
(b) Each prelexical utterance was coded for its structure as a sequence of consonant- and vowel-like elements. This coding permits the automatic selection of CV syllables from the transcripts.
(c) Each consonant-like element was coded for various articulatory features. This coding permits us to automatically identify true consonants as required for the TCBR measures.

Visual information available from the video images guided decisions on segmental characteristics.

All transcribed prelexical utterances were subsequently syllabified with an automated procedure based on universal principles for syllabification (viz. the CORE SYLLABIFICATION PRINCIPLE; Clements, 1990). The syllabified materials were used for determining the number of canonical, non-canonical, and the total number of syllables for the calculation of the CBR and TCBR ratios.

Table 1 provides an overview of the data. For each child the numbers of prelexical utterances and syllables are provided in the samples from age 0;6 up to 1;0.

Reliability

An extensive part of the large speech corpus underwent a reliability check to assess both between- and within-transcriber agreement. For the scoring of between-transcriber reliability, 10% of the original transcriptions in the

TABLE 1. *Overview of the numbers of utterances* (*Utts*) *and syllables* (*Syls*) *in the samples between* 0;6 *and* 1;0 *for each of the participants in the study* (*slashes indicate missing sessions*)

	0;6		0;7		0;8		0;9		0;10		0;11		1;0	
	Utts	Syls	Utts	Syls	Utts	Syls	Utts	Syls	Utts	Syls	Utts	Syls	Utts	Syls
P1	149	486	313	492	210	267	237	340	181	348	272	434	226	346
P2	296	319	362	470	316	354	396	515	289	461	236	472	171	305
P3	330	620	345	585	239	455	323	514	264	488	214	418	322	593
P4	225	243	377	490	332	420	338	668	280	497	280	507	371	567
P5	199	229	328	618	210	282	150	226	257	464	273	511	227	425
P6	169	256	193	279	135	184	201	347	151	252	169	272	243	374
P7	143	301	146	258	336	506	244	409	294	512	174	309	203	314
P8	180	271	203	267	173	290	291	490	190	285	221	345	230	305
P9	134	167	247	387	353	551	256	397	261	482	304	397	178	370
P10	113	305	233	474	184	289	234	331	282	438	279	472	164	219
P11	210	273	251	338	204	248	182	238	138	185	113	145	252	290
P12	270	474	179	326	335	693	365	848	287	533	265	587	196	405
P13	290	517	209	289	143	210	178	349	199	430	209	333	213	325
P14	98	122	109	143	118	227	219	342	234	328	229	345	137	200
P15	265	300	204	316	414	645	325	469	336	671	263	397	304	488
P16	354	395	312	431	279	404	309	447	272	428	183	371	232	519
P17	247	369	247	417	238	340	207	359	255	494	181	359	176	266
P18	229	277	287	571	206	288	267	364	247	419	271	380	274	490
P19	153	262	168	265	245	392	157	289	202	405	183	390	157	271
P20	392	417	242	336	307	521	349	675	306	458	244	449	230	447
P21	237	281	131	148	326	397	358	492	319	412	258	310	280	331
P22	149	235	166	315	189	340	273	437	288	527	161	305	179	261
P23	166	283	133	194	238	414	164	472	156	379	193	475	121	258
P24	202	280	244	348	238	309	323	655	231	388	190	404	250	617
P25	220	489	538	882	282	368	298	470	167	319	168	383	134	262
P26	190	407	167	333	224	401	181	356	192	428	224	447	211	416
P27	382	561	316	425	144	259	263	345	242	378	320	487	293	465
P28	99	159	271	336	197	278	186	325	162	319	211	332	271	397
P29	231	290	214	306	198	253	192	270	79	127	157	304	172	417
P30	295	505	272	468	470	661	221	516	409	632	289	718	314	717
P31	345	444	451	605	561	831	394	587	371	610	276	471	144	228
P32	/	/	/	/	104	135	177	273	119	158	193	297	161	219
P33	380	588	265	403	401	526	345	667	265	378	204	323	/	/
P34	/	/	/	/	312	490	279	538	/	/	/	/	317	579
P35	305	572	270	349	160	230	263	398	214	408	346	579	198	350
P36	197	254	362	625	386	730	311	613	320	562	311	538	81	135
P37	/	/	165	219	231	323	362	470	208	309	246	402	366	540
P38	366	527	293	400	247	349	272	458	310	530	321	540	202	292
P39	423	550	334	491	236	348	240	347	241	346	107	194	244	424
P40	299	363	213	263	170	215	269	418	420	531	238	367	268	361

database was retranscribed by a second, equally experienced transcriber who remained blind to the original transcription. For the coding of prelexical vocalizations relevant to the present paper, this procedure yielded the following reliability scores: for the decision if utterances contained

multiple articulatory movements combined with interrupted or uninterrupted phonation, Cohen's kappa = 0·65; for the coding in consonant-like and vowel-like segments, Cohen's kappa = 0·58; and for the coding of manner of articulation of consonant-like segments as 'true' consonants as opposed to 'other' consonants, Cohen's kappa = 0·52. All kappa scores for between-transcriber agreement fall within the range of 'moderate' agreement, according to the evaluation by Landis and Koch (1977).

For the assessment of within-transcriber agreement, the original transcriber fully retranscribed 5% of the material after a pause of at least two months. Within-transcriber reliability scores were: for the coding of utterances in terms of phonation and articulation, Cohen's kappa = 0·75; for the coding of consonant-like and vowel-like segments, Cohen's kappa = 0·76; and for the coding of manner of articulation of consonant-like segments as 'true' versus 'other' consonants, Cohen's kappa = 0·67. All kappa scores for within-transcriber agreement are indicative of 'substantial' agreement.

Bootstrap

In order to establish the onset of babbling in a reliable way for all children in the study, a bootstrapping procedure was implemented (Baayen, 2008; Efron, 1979). The aim of the procedure was to establish a single sample size that can be used as a general guideline for reliably establishing the onset of babbling. The procedure consists of extensively resampling and is written in pseudo-code in (5).

(5)
```
For each CHILD
    BabblingOnset:=UNDEFINED;
    AGE:=0;6;
    Repeat
        SampleSize:=0;
        Repeat
            SampleSize +=25;
            For i:=1 to 1,000
                Select a random sample (RS);
                ComputeMeasure (e.g. CBR^syl) for RS;
                If CBR^syl for RS >=0·15 then AboveThreshold +=1;
            If AboveThreshold >=950/1,000
                Then BabblingOnset:=AGE;
                Minimal_Sample_Size:=SampleSize
        Until condition (7) holds
        If BabblingOnset=UNDEFINED
            AGE +=0;1
Until BabblingOnset=DEFINED
```

The bootstrapping procedure is used to establish the age at which each child for the first time reaches the babbling threshold, for instance $CBR^{syl} >= 0{\cdot}15$ (see (2)). This is accomplished by taking for the first age for which data are available (0;6 in the present study) a random sample of 25 syllables from the transcript and computing the CBR^{syl} for that sample. The random sampling (after replacement) of 25 syllables is repeated 1,000 times, and each time CBR^{syl} is computed, so that after 1,000 iterations the proportion of samples that are equal to or above the threshold is established. If this threshold is not reached in 95% of the cases with samples of 25 syllables, the sample size is incremented with 25 syllables, and the procedure is repeated for 50 syllables, and so on. The age at which the child starts babbling is reached when in 95% (or more) of the samples the metric at least equals the preset threshold. If the data for a particular age (month) do not permit to establish the onset of babbling reliably, i.e. for no sample size a confidence level of 95% can be reached, the data for the next month are entered into the procedure.

The initial sample size of 25 items (syllables or utterances, depending on the measure) was set arbitrarily, as well as the increment of 25 items. The maximum sample size, i.e. the ceiling of the sampling procedure, was determined in a principled way: the number of combinations of size k from a set of size n was computed according to equation (6), the binomial coefficient:

$$C(n,k) = \frac{n!}{k!(n-k)!} \tag{6}$$

This means, for instance, that for the smallest session (subject P29 at 0;10: 79 utterances; see Table 1), a sample size of 75 items still leaves 1,502,501 unique (though (partially) overlapping) combinations to select from. Each of the 1,000 iterations can thus in principle be run with a different sample of 75 items. Hence, the sample size was increased under the conditions in (7):

(7) k <n and C(n,k) <r
with k=the sample size, n=the number of utterances/syllables (depending on the measure) in a session, and r=the number of runs (i.c. 1,000)

Thus for subject P29 mentioned above, samples sizes of 25, 50 and 75 were entered in the bootstrapping procedure: $n=79$ and $k=75$, so that $k < n$.

In each run of the bootstrap, the age at babbling onset is computed according to each of five criteria:

(a) Canonical Babbling Ratio $>0{\cdot}20$ (CBR^{utt}): babbling onset is defined as the first age at which in at least 950 out of 1,000 runs of the bootstrapping procedure the number of canonical syllables (syllables

containing one vowel-like element and at least one non-glottal consonant) divided by the total number of utterances equals or exceeds 0·20 (see (1));

(b) CANONICAL BABBLING RATIO >0·15 (CBR^{syl}): babbling onset is defined as the first age at which in at least 950 out of 1,000 runs the number of canonical syllables (syllables containing one vowel-like element and at least one non-glottal consonant) divided by the total number of syllables equals or exceeds 0·15 (see (2));

(c) TRUE CANONICAL BABBLING RATIO >0·20 ($TCBR^{utt}$): babbling onset is defined as the first age at which in at least 950 out of 1,000 runs the number of canonical syllables (syllables containing one vowel-like element and at least one non-glottal, non-glide consonant) divided by the total number of utterances equals or exceeds 0·20 (see (3));

(d) TRUE CANONICAL BABBLING RATIO >0·15 ($TCBR^{syl}$): babbling onset is defined as the first age at which in at least 950 out of 1,000 runs the number of canonical syllables (syllables containing one vowel-like element and at least one non-glottal, non-glide consonant) divided by the total number of syllables equals or exceeds 0·15 (see (4));

(e) MULTISYLLABICITY REQUIREMENT (MULTI): babbling onset is defined as the first age at which in at least 950 runs out of 1,000 two or more utterances combining continuous or interrupted phonation with multiple articulatory movements are present, provided the number of such babbled utterances does not decrease to less than two in samples from the two following months.

RESULTS

Repeatedly computing the measures

In the literature the various measures for the identification of babbling onset ((T)$CBR^{utt/syl}$ and MULTI) are computed on one single speech sample for each age. For instance, Oller and Eilers (1988) propose a sample size of 50 utterances to compute CBR^{utt}. In Table 2 the results are shown of computing CBR^{utt} 1,000 times using randomly drawn samples of 50 utterances. For this simulation the data of child P40 are used from 0;6–1;0.

The results in Table 2 show that the mean value of CBR^{utt} at 0;9 is larger than 0·20 (the threshold for babbling onset). However, at that age the CBR^{utt} values range from 0·06 to 0·60, which means that if only one sample had been taken to determine the onset of babbling, CBR^{utt} could have pointed in either direction: the child could or could not have been credited with onset of babbling. Depending on the hazards of selecting the speech sample, the child could have started babbling at age 0;8 – the age at which the maximum value of CBR^{utt} exceeds 0·20 for the first time — or the onset of babbling could not have been reached at 1;0: the minimum values up to

TABLE 2. *CBR^{utt} computed on 1,000 randomly drawn samples of 50 utterances from 0;6–1;0 of child P40*

	0;6	0;7	0;8	0;9	0;10	0;11	1;0
Mean	0·06	0·05	0·10	0·30	0·10	0·46	0·34
Median	0·06	0·04	0·10	0·30	0·10	0·46	0·32
Minimum	0·00	0·00	0·00	0·06	0·00	0·14	0·04
Maximum	0·16	0·16	0·26	0·60	0·26	0·80	0·72

TABLE 3. *Overview of the ages at onset of babbling and the required minimal sample sizes using bootstrapping procedure* (5)

Measure	Onset babbling (median, age range)	Sample size (range)
CBR^{syl}	0;7 (0;6−0;9)	25−425
$TCBR^{syl}$	0;7·5 (0;6−0;11)	25−500
CBR^{utt}	0;7 (0;6−0;9)	25−375
$TCBR^{utt}$	0;7 (0;6−0;11)	25−300
MULTI	0;6 (0;6−0;11)	50−350

age 1;0 are all below 0·20. Similar results hold for the other babbling measures. Hence it does not suffice to compute a measure on a single isolated sample in order to arrive at a reliable decision as to the onset of babbling. In other words, it is necessary to repeatedly compute the various measures so that the variance in the data can be estimated.

Results of bootstrapping

The results of the bootstrapping procedure using the algorithm outlined in (5) are displayed in Table 3. For each measure, Table 3 shows the median age at onset of babbling and the sample size needed to reach 95% confidence in the result. The number of items (syllables or utterances) required for reliably establishing babbling onset depends on the child and on the particular measure. The amounts range from a minimum of 25 to 300 utterances for $TCBR^{utt}$, and maximally 350 utterances for MULTI, 375 utterances for CBR^{utt}, 425 syllables for CBR^{syl} and 500 syllables for $TCBR^{syl}$. This means that, based on the results for the forty children in our study group, for $TCBR^{syl}$ samples of at least 500 syllables are needed in order to decide with 95% confidence that a child has started babbling using only a single computation of the measure.[1]

[1] These figures are determined empirically. But probability theory gives similar indicative figures. Given a phenomenon that occurs in a particular proportion of the data (e.g.

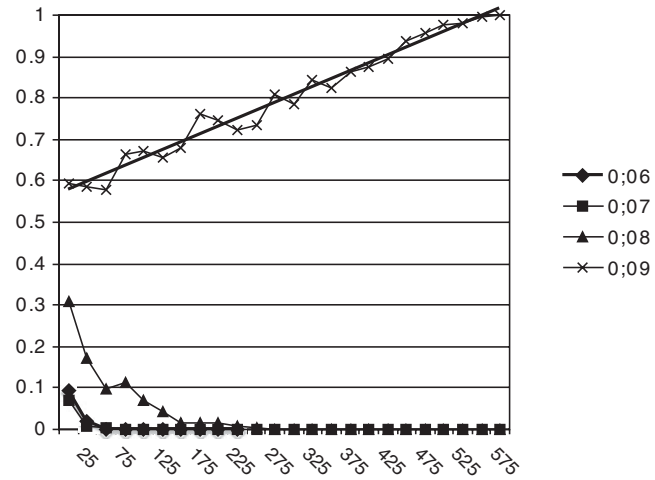

Fig. 1. Proportion of samples at or above the critical threshold of $TCBR^{syl}=0{\cdot}15$ for sample sizes 25–600 syllables for child P36.

Effect of sample size

In Figure 1, the results of the bootstrapping procedure outlined in (5) are illustrated for the computation of $TCBR^{syl}$ with sample sizes increasing from 25 syllables up to 600 syllables for child P36 from 0;6 up to the onset of babbling.

The graph shows that at 0;6 a sample size of 25 utterances yields only 93 out of 1,000 $TCBR^{syl}$ values at or above the critical threshold for babbling onset which was set at 0·15. That number decreases with an increase of the sample size. The child can thus not be credited with babbling onset at age 0;6. At 0;7 and 0;8 the threshold for babbling onset is not reached in enough samples either. At 0;8, for instance, there are 309 samples of 25 syllables reaching the threshold. That number decreases to 173 for 50 syllables, 93 for 75 syllables and further down to 0 from 300 syllables onwards. At 0;9, 593 out of 1,000 samples of 25 syllables are at or above $TCBR^{syl}=0{\cdot}15$. However, with increasing sample size the 95% confidence level is eventually reached, viz. when samples of 500 syllables are used. Hence in the case of P36 the onset of babbling according to the $TCBR^{syl}$ criterion can be set at age 0;9 (for a sample size of 500 syllables).

As Figure 1 illustrates, by applying the measure on samples of increasing size, a curve can be drawn that exhibits a clear slope. This slope indicates

babbling in 20% of the cases, for $(T)CBR^{utt}$), the required minimal sample size can be computed. With 95% confidence and a precision of 0·05, 384 utterances are needed for CBR^{utt}, and if the required precision is brought to 0·01, 9,513 utterances are required (Woods, Fletcher & Hughes, 1986).

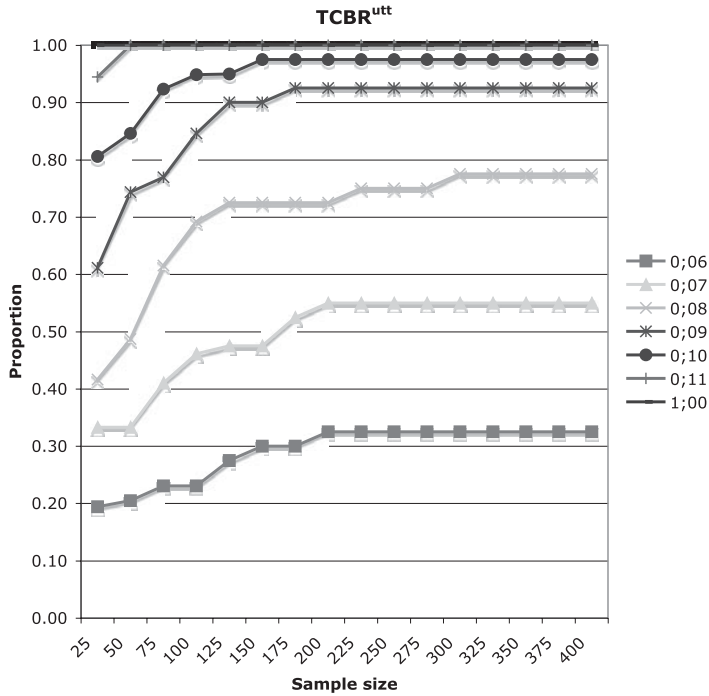

Fig. 2. Cumulative proportion of children reaching the babbling onset threshold for $TCBR^{utt}$.

unambiguously whether the measure is coming closer to and will eventually reach the threshold or not. But a sufficient number of data points is needed in order to fit the data well: at 0;9, the first three data points show decreasing values (a negative slope), while eventually, the slope of a linear fit is positive. Suppose the data for P36 at 0;9 only permitted samples of 75 syllables. Less than 95% of the samples would have reached criterion, and onset of babbling would not have been credited. Or, if in that case a regression was applied, the direction of the slope would have been negative, with the same result: no babbling onset at 0;9.

The effect of the size of the sample is illustrated more generally in Figure 2, which depicts for $TCBR^{utt}$ per age the cumulative number of children that start babbling. For reasons of clarity, the graph is restricted to samples of 25 to 400 utterances. At 0;6, 20% of the children have started babbling, when samples of 25 utterances are considered. That number increases to 23% when samples of 75 or 100 utterances are drawn, and it further increases until samples of 200 utterances are drawn. After that point, no further increase is noted, so that it can be concluded that for

$TCBR^{utt}$ at age 0;6 at least 200 utterances are needed in order to reliably compute the onset of babbling. At 0;7 33% of the children reach the babbling threshold with samples of 25 utterances, but that number increases to over 50% by the time samples of 200 utterances are used in the procedure. At 0;8 it even takes samples of 300 utterances to reach the maximal proportion of children who can be considered to babble according to $TCBR^{utt}$. Thus, we can conclude that for $TCBR^{utt}$ at least 300 syllables are needed in order to compute the measure reliably at 0;8. In more general terms: the larger the sample, the more children reach the critical threshold that marks onset of babbling at any given age. This was true for each of the five measures under investigation.

For none of the measures was a significant correlation found between the age at onset of babbling and the required sample size (Spearman's rho: CBR^{syl}: rho = 0·1, $p = 0·33$; CBR^{utt}: rho = 0·26, $p = 0·10$; $TCBR^{syl}$: rho = −0·16, $p = 0·32$; $TCBR^{utt}$: rho = −0·002, $p = 0·99$; MULTI: rho = −0·20, $p = 0·23$). Hence, there is no clear-cut relationship between the age at onset of babbling and the size of the sample required to compute it.

The results in Figure 2 imply that an equal amount of data (syllables or utterances) should be considered for each subject and each age. The bootstrap procedure shows that otherwise this may lead to the attribution of babbling onset to a particular child and not to another child solely because of a difference in the amount of data available. Both could have been credited for babbling if an equal amount of data had been considered.

DISCUSSION

The aim of the first experiment was to establish a minimally required sample size for reliably computing the onset of babbling in one single pass given a relatively limited recording of a sample of spontaneous speech. In other words, the question addressed was: How large should a sample be so that the computation of $(T)CBR^{syl/utt}$ or MULTI yields a reliable indication of whether a child has started babbling or not?

A first finding was that in order to establish the minimal sample size the measures should be applied repeatedly on the data so that the variance can be measured. A bootstrapping experiment was designed implementing a specific algorithm for repeatedly measuring $(T)CBR^{syl/utt}$ and MULTI for increasing sample sizes. The main result of the experiment was that if a single computation of the measures is aimed at, and a confidence of 95%, reasonably large samples are required: at least 300 utterances for $TCBR^{utt}$ and at least 500 syllables for $TCBR^{syl}$, for example (see Table 3). In order to put these findings into practice, these sample sizes are required for all children, otherwise erroneous conclusions with respect to babbling onset can be drawn in individual cases.

This poses a serious problem: inspection of Table 1 reveals that, especially at the younger ages, such amounts of utterances/syllables are not available in the transcriptions. Are there solutions for this impasse?

A first solution is rather pragmatic in nature: corpora such as the one used in this study are rather limited. That is, each recording consists of a limited number of data (utterances, syllables). If an equal number of utterances or syllables is required for further analysis, the size of the samples can be determined by the recording with the least data. A quick overview of Table 1 reveals that the smallest number of utterances and syllables is produced by P19 at 0;10: 79 utterances and 127 syllables. Thus the smallest sample can be used as benchmark, taking into account the fact that in so doing the number of children with babbling onset in a particular month will be (vastly) underestimated.

A second solution will be explored in Experiment 2: the quest for finding a minimal sample size for which a single computation of the five measures suffices is abandoned. Instead, a 'lenient' bootstrapping procedure – as opposed to the strict version in (5) – is introduced, which permits establishing a minimal sample size that is sufficient to identify the onset of babbling given that the (T)CBR$^{syl/utt}$ and the MULTI measures are computed in 1,000 runs in a bootstrapping procedure.

EXPERIMENT 2

METHOD

Data

The same data as in Experiment 1 were used.

Bootstrap

The bootstrapping procedure described in (5) is a stringent one: it aims at the identification of a sample size that permits computing the babbling measures with 95% confidence using one single pass over a transcript. The second bootstrapping procedure that was implemented is more lenient. It makes use of the idea that by iteratively selecting a random sample, a good estimate of the sample's distribution is discovered. The algorithm is expressed in pseudo-code in (8).

(8)
```
For each CHILD
    BabblingOnset:=UNDEFINED;
    AGE:=0;6;
    Repeat
        SampleSize:=0;
        Repeat
```

```
        SampleSize + = 25;
        For i: = 1 to 1,000
            Select a random sample (RS);
            ComputeMeasure (e.g., CBR^syl) for RS;
        Compute Mean_for_SampleSize;
        Compute Mean_for_SampleSize Upper 95%;
        Compute Mean_for_SampleSize Lower 95%;
    Until condition (7) holds;
    SampleSize: = 25;
    Repeat
    If (Mean_for_SampleSize_i > threshold) and
        (Mean_for_SampleSize_i Upper 95% > threshold) and
        (Mean_for_SampleSize_i Lower 95% > threshold)
        Then  BabblingOnset: = AGE
            Minimal_Sample_Size: = SampleSize_i
        Else i + = 25
    Until BabblingOnset = DEFINED or all data for AGE are used
    If BabblingOnset = UNDEFINED then Age + = 0; 1;
Until BabblingOnset = DEFINED
```

The algorithm in (8) proceeds in two steps. In the first step, the measure (e.g. CBR^{syl}) is computed repeatedly for increasing sample sizes (the limit is defined by the conditions mentioned in (7)), and for each sample size the mean value is determined, as well as the mean for the upper 95% of the values and the mean of the lower 95% of the values. In the second step, the algorithm determines at what age and with which minimal sample size the onset of babbling can be established with reasonable confidence.

Thus, the first step is an iterative process: for each age the value of a measure, e.g. CBR^{utt}, is computed 1,000 times starting with random samples of 25 items, and increasing the sample size with 25 items until the conditions in (7) hold. For each batch of 1,000 runs the mean value is computed, so that a mean CBR^{utt} value results for the various sample sizes (M_{25}, M_{50}, M_{75}, ...), together with the mean of the upper 95% and the lower 95% of the values. The second step consists of comparing those mean values with the threshold (which is 0·20 for CBR^{utt}), starting with the smallest sample size (M_{25}). If M_{25} passes the threshold (e.g. $CBR^{utt} > = 0{\cdot}20$), and the means of the upper and lower 95% of the values do the same, then the age at babbling onset is established, and the minimal sample size for babbling onset is established: $N = 25$. If the means over 1,000 runs do not pass the threshold, then the onset of babbling is not credited yet, and the result for a bigger sample size (M_{50}) is investigated, or eventually the result for the following age. The evaluation of the upper 95% and the lower 95% mean come into play especially in cases where the mean value is very close

to the threshold: if both are at the same side of the threshold (above or below), then the appropriate consequence is followed (see above). If the upper 95% mean is above, but the lower 95% mean is below the threshold, then the sample size is increased with 25 items.

This bootstrapping procedure has a different aim from the first one. With the bootstrapping procedure outlined in (5), the aim was to find the minimal sample size that allows us to discover babbling onset with 95% confidence using one single pass over a dataset. The bootstrapping procedure described in (8) is meant to identify for each babbling measure a sample size that can be used to reliably decide when a child starts babbling, provided that the measure is computed 1,000 times. The procedure implements the idea that by iteratively selecting a random sample, a good estimate of the sample's distribution is discovered, and hence the sample's 'real' mean is approximated. The bootstrapping procedures will be further exemplified in the 'Results' section.

RESULTS

Repeatedly computing the measures

The 'lenient' bootstrapping procedure outlined in (8) requires the repeated computation of a measure, e.g. $TCBR^{syl}$, for increasing sample sizes. A typical result of 1,000 runs of the $TCBR^{syl}$ for increasing sample sizes is displayed in Figure 3a. This shows the results for subject P30 at 0;6. In computing the minimally required sample size, the following steps are taken: for each sample size the mean $TCBR^{syl}$ value is computed over 1,000 runs, as well as the 95% upper mean and the 95% lower mean. If all mean values for all sample sizes point in the same direction for the onset of babbling, then the smallest sample size suffices, since onset of babbling is a binary decision. In this example the $TCBR^{syl}$ values centre around the mean across all sample sizes, and the mean $TCBR^{syl}$ value for all sample sizes is well above the critical threshold, i.e. $TCBR^{syl}=0{\cdot}15$. Thus, even for a sample size of 25 syllables, the mean value of 0·2309 (as well as the lower and upper 95% means, 0·2258 and 0·2360, respectively) is well over the threshold value. And thus we can safely assume that P30 has started babbling by 0;6. Hence, for P30 a sample size of 25 syllables can be retained as the minimally required sample size (see Table 4).

Figure 3b shows an example of a problematic case. The mean $TCBR^{syl}$ value for subject P32 at 0;9 is close to the threshold ($TCBR^{syl}=0{\cdot}15$): with samples of 25 syllables the mean value is 0·1489. In order to decide in this and similar cases on the onset of babbling, the algorithm in (8) contains extra conditions, viz. if the lower 95% mean as well as the upper 95% mean are above the threshold then the child is said to have reached babbling onset. But in this case the lower 95% mean equals 0·1489, which is below

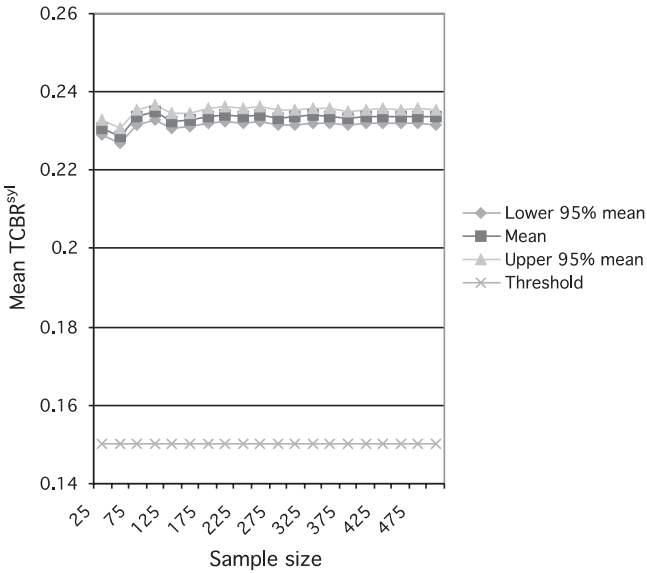

Fig. 3a. Mean $TCBR^{syl}$ values for subject P30 at 0;6 for increasing sample sizes and 1,000 random samples for each sample size.

the threshold, and the upper 95% mean equals 0·1533, which is above the threshold. In that case, the sample size is incremented by 25 syllables. In the example of P32 for $N = 50$, the lower 95% mean equals 0·1423, and the upper 95% mean equals 0·1479, both are below the threshold. Figure 3b clearly shows that incrementing the sample size does not bring a change: the mean $TCBR^{syl}$ values remain below the critical threshold, and thus P32 cannot be said to start babbling at 0;9.

Results of bootstrapping

Table 4 provides the relevant results of the bootstrapping exercise for the forty children in this study for the five measures. The table shows for every child the age at onset of babbling according to each of the five measures. In addition the table reveals how many items (syllables or utterances) were required to attain that decision. These data can now be used to define the minimal sample size required to reliably compute the onset of babbling, by taking for each measure the largest value in the column 'Sample size'. Based on the present data from forty children, the following sample sizes are required for reliably computing the five measures, given a bootstrapping procedure with 1,000 iterations: samples of 25 syllables suffice for the $(T)CBR^{utt/syl}$ measures, while 75 utterances are required for MULTI.

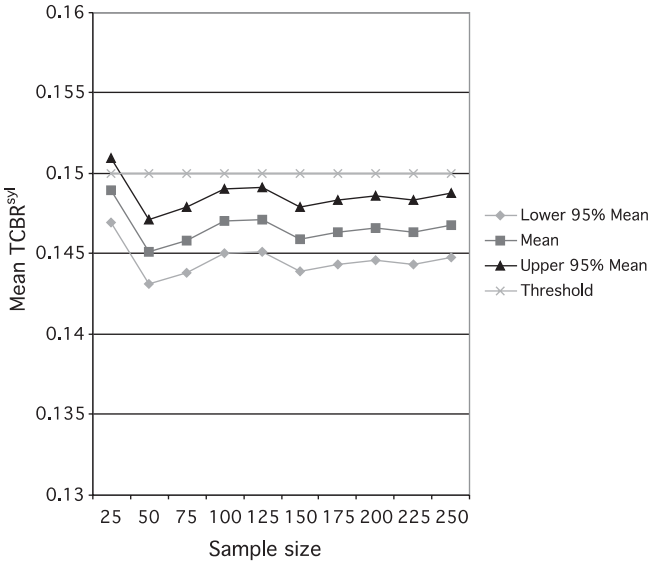

Fig. 3b. Mean $TCBR^{syl}$ values for subject P32 at 0;9 for increasing sample sizes and 1,000 random samples for each sample size.

Comparison of the outcomes

How comparable are the outcomes of the five measures? Does the onset of babbling for a particular child differ considerably depending on the measure that is applied? Or do the five measures basically measure the same phenomenon without much variation, i.e. is the age at onset of babbling for a particular child the same irrespective of the measure that is applied? With the results of the second experiment, i.e. reliable ages at babbling onset for all children in the cohort based on a 'lenient' bootstrap procedure, this question can be answered in three ways. First of all, correlations are computed between the ages at onset of babbling. In Table 5 the Spearman correlations are displayed between the various measures (computed given the sample sizes decided on in the previous paragraph). Highly significant correlations hold between the four (T)CBR measures. MULTI also correlates significantly with the (T)CBR measures, except for $TCBR^{syl}$, which is not significantly correlated with MULTI.

Second, pairwise comparisons can be made of the age at babbling onset as determined by the various measures. For instance, for how many children do CBR^{syl} and CBR^{utt} compute the same month as babbling onset? The matrix in Table 6 contains that information for each pair of measures. It appears from the table that CBR^{syl} and CBR^{utt} compute the same month for babbling onset for 36 out of 40 children (90%). $TCBR^{syl}$ and $TCBR^{utt}$ make

TABLE 4. *Age at onset of babbling and sample size using bootstrapping procedure* (8) *for forty children*

	CBR^{syl}		$TCBR^{syl}$		CBR^{utt}		$TCBR^{utt}$		Multisyllabicity	
	Age onset	Sample size	Age onset	Sample size	Age onset	Sample size	Age onset	Sample size	Age onset	Sample size
P1	0;7	25	0;7	25	0;7	25	0;7	25	0;10	25
P2	0;9	25	0;10	25	0;9	25	0;10	25	0;10	25
P3	0;6	25	0;6	25	0;6	25	0;6	25	0;6	25
P4	0;7	25	0;8	25	0;8	25	0;8	25	0;9	25
P5	0;7	25	0;7	25	0;7	25	0;7	25	0;10	25
P6	0;8	25	0;9	25	0;6	25	0;9	25	0;9	25
P7	0;6	25	0;9	25	0;6	25	0;9	25	0;6	25
P8	0;6	25	0;6	25	0;6	25	0;6	25	0;8	25
P9	0;7	25	0;8	25	0;7	25	0;8	25	0;8	25
P10	0;7	25	0;7	25	0;7	25	0;7	25	0;7	25
P11	0;7	25	0;7	25	0;7	25	0;7	25	0;9	75
P12	0;6	25	0;7	25	0;6	25	0;6	25	0;6	25
P13	0;6	25	0;9	25	0;6	25	0;9	25	0;9	25
P14	0;7	25	0;7	25	0;7	25	0;7	25	0;11	25
P15	0;7	25	0;8	25	0;7	25	0;7	25	0;10	25
P16	0;7	25	0;8	25	0;7	25	0;8	25	0;10	25
P17	0;7	25	0;8	25	0;7	25	0;8	25	0;9	25
P18	0;6	25	0;7	25	0;7	25	0;7	25	0;9	25
P19	0;6	25	0;6	25	0;6	25	0;6	25	0;8	25
P20	0;8	25	0;9	25	0;8	25	0;9	25	0;8	25
P21	0;6	25	0;6	25	0;6	25	0;6	25	0;7	75
P22	0;7	25	0;8	25	0;7	25	0;8	25	0;6	25
P23	0;6	25	0;6	25	0;6	25	0;6	25	0;7	25
P24	0;6	25	0;8	25	0;6	25	0;7	25	0;7	25
P25	0;6	25	0;6	25	0;6	25	0;6	25	0;10	25
P26	0;6	25	0;6	25	0;6	25	0;6	25	0;6	25
P27	0;6	25	0;6	25	0;6	25	0;6	25	0;10	25
P28	0;6	25	0;6	25	0;6	25	0;6	25	0;9	25
P29	0;7	25	0;7	25	0;7	25	0;7	25	0;9	50
P30	0;6	25	0;6	25	0;6	25	0;6	25	0;6	25
P31	0;8	25	0;8	25	0;8	25	0;8	25	0;11	25
P32	0;9	25	0;11	25	0;9	25	0;9	25	0;11	25
P33	0;6	25	0;10	25	0;6	25	0;10	25	0;9	25
P34	0;8	25	0;8	25	0;8	25	0;8	25	/	
P35	0;6	25	0;6	25	0;6	25	0;6	25	0;10	25
P36	0;9	25	0;9	25	0;8	25	0;8	25	0;9	25
P37	0;7	25	0;8	25	0;7	25	0;8	25	0;10	25
P38	0;6	25	0;6	25	0;6	25	0;6	25	0;8	25
P39	0;7	25	0;7	25	0;7	25	0;7	25	0;6	25
P40	0;9	25	0;11	25	0;9	25	0;11	25	0;11	50
Min	0;6	25	0;6	25	0;6	25	0;6	25	0;6	25
Max	0;9	25	0;11	25	0;9	25	0;11	25	0;11	75
Median	0;7	25	0;7	25	0;7	25	0;7	25	0;7	25

exactly the same decision in 31 out of 40 cases (78%). The least overlap is found between $TCBR^{syl}$ and MULTI: only in 14 out of 39 cases (36%) do both measures arrive at exactly the same month.

TABLE 5. *Spearman ρ correlations between the ages of babbling onset determined by the five measures*

	CBR^{syl}	$TCBR^{syl}$	CBR^{utt}	$TCBR^{utt}$	MULTI
CBR^{syl}		$\rho = 0{\cdot}67$ $p = 0{\cdot}0001$	$\rho = 0{\cdot}89$ $p < 0{\cdot}0001$	$\rho = 0{\cdot}77$ $p < 0{\cdot}0001$	$\rho = 0{\cdot}40$ $p = 0{\cdot}01$
$TCBR^{syl}$			$\rho = 0{\cdot}60$ $p = 0{\cdot}0001$	$\rho = 0{\cdot}89$ $p < 0{\cdot}0001$	$\rho = 0{\cdot}26$ $p = 0{\cdot}11$
CBR^{utt}				$\rho = 0{\cdot}69$ $p < 0{\cdot}0001$	$\rho = 0{\cdot}38$ $p = 0{\cdot}02$
$TCBR^{utt}$					$\rho = 0{\cdot}38$ $p = 0{\cdot}02$
MULTI					

TABLE 6. *Agreement between the five measures as to onset of babbling expressed in number of children* (*figures in parentheses : the median difference in number of months and the range*)

	CBR^{syl}	$TCBR^{syl}$	CBR^{utt}	$TCBR^{utt}$	MULTI
CBR^{syl}		22 (0, 0−4)	36 (0, 0−2)	26 (0, 0−3)	23 (0, 0−3)
$TCBR^{syl}$			23 (0, 0−4)	31 (0, 0−3)	14 (1, 0−3)
CBR^{utt}				27 (0, 0−3)	22 (0, 0−3)
$TCBR^{utt}$					19 (1, 0−3)
MULTI					

Third, for each measure a cumulative count of the children that are credited with the onset of babbling in each month is displayed in Figure 4. The graph in this figure confirms the significant correlations reported in Table 5: in general it can be observed that although the slopes of the curves differ, there is a large isomorphism. Two additional observations can be made: CBR^{syl} and CBR^{utt} hardly differ in determining the onset of babbling, and $TCBR^{syl}$ appears to be the most conservative measure.

In sum, Figure 4 indicates that all children in the cohort can be credited for babbling onset between the ages of 0;6 and 0;11 with each of the measures (note that there is a flooring effect because our data collection only started at the age of 0;6). Measured by CBR^{syl} and CBR^{utt}, the age range for babbling onset is even 0;6–0;9. For all measures, the median age at babbling onset for the forty typically developing children in our study group lies between 0;6 and 0;7.

DISCUSSION

An outcome of the first experiment was that in order to reliably compute the onset of babbling, relatively large samples were needed (300 or more

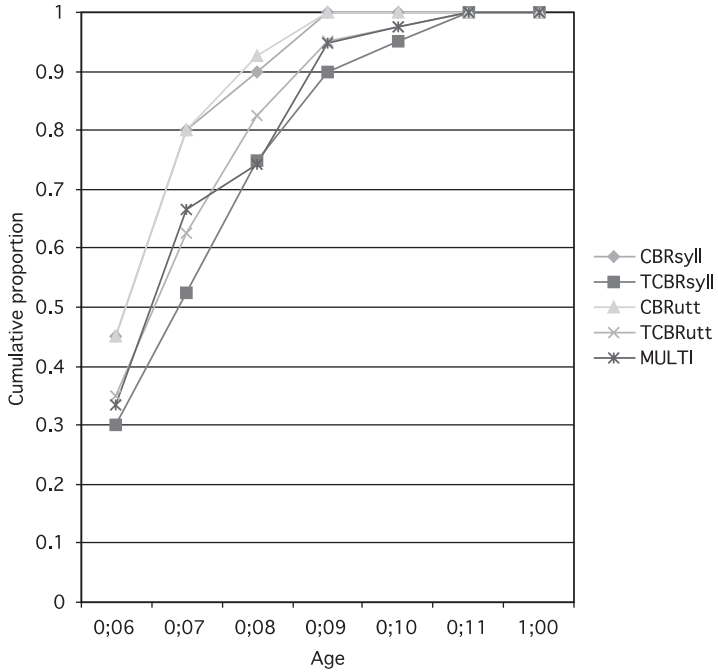

Fig. 4. Cumulative proportions of children credited for onset of babbling by CBR^{syl}, CBR^{utt}, $TCBR^{syl}$, $TCBR^{utt}$ and MULTI.

syllables or utterances) and that the same amount of data was required for all children. Since spontaneous speech samples of very young children tend to be rather small (notwithstanding a huge investment of recording and transcription time) and since the volubility of young children is rather variable (Molemans, Van Severen, van den Berg, Govaerts & Gillis, 2010), a new, more lenient bootstrapping procedure was developed and tested in the second experiment. The aim of that procedure was to establish for each measure the size that subsamples should have in order to reliably determine the onset of babbling. But instead of looking for a sample size that enables reaching that goal with one computation of the measure, an iterative process of taking 1,000 random samples of a particular size is implemented. This procedure yields 1,000 values for the measure, and the mean value is thought to approximate the 'real' value for that measure quite reliably.

The result of the experiment on the data from forty children provides sample sizes for the five measures under investigation. For $(T)CBR^{utt/syl}$, samples of 25 syllables/utterances are required. This means that in order to construct 1,000 unique random samples, a transcription with at least 28

different syllables/utterances should be available according to equation (6). For MULTI, 75 utterances are required, which implies that there should be at least 77 utterances in the transcript.

How much do the various measures differ in establishing the onset of babbling? There are highly significant correlations between the ages at which the various measures establish the onset of babbling (Table 5). The lowest – but still significant, except for one comparison – correlations are between MULTI and the other measures. In quite a number of cases, the various measures result in exactly the same month of babbling onset: for ten children the five measures make the same evaluation, for another thirteen children four measures yield the same result, and in an additional twelve children three measures yield the same month. However, this should not lead to the conclusion that the five measures can be used interchangeably: as shown in Table 6, the median difference between each pair of measures is zero months for most comparisons, but the range is from zero to up to four months. This means that depending upon the particular measure used, a child can be said to start babbling up to four months earlier or later. Thus, the choice of the measure to determine onset of babbling influences the result to a certain extent, and the different measures do not measure exactly the same phenomenon.

GENERAL DISCUSSION AND CONCLUSION

We set out to investigate five measures for determining the onset of babbling that were introduced in the literature, viz. $CBR^{utt/syl}$, $TCBR^{utt/syl}$ and MULTI. These measures differ in several respects. First of all, CBR and TCBR measure the number of canonical syllables relative to a number of utterances (CBR^{utt} and $TCBR^{utt}$) or relative to a number of syllables (CBR^{syl} and $TCBR^{syl}$). Second, in calculating $CBR^{syl/utt}$, all canonical syllables with supraglottal consonants are taken into account, while in calculating $TCBR^{syl/utt}$, syllables with glides as margins are left out of the procedure, leaving only the 'true' supraglottal consonants (Stoel-Gammon, 1989). Third, while the $(T)CBR^{utt/syl}$ measures compute the ratio of canonical syllables on the total number of syllables or utterances, MULTI takes the repetition of a syllabic structure into account. In the terminology of van der Stelt & Koopmans-van Beinum (1986), babbling requires (un)interrupted phonation with several (two or more) articulatory movements. In addition, MULTI also incorporates a longitudinal perspective: a minimal number of babbles is required in three consecutive months (observation sessions). We set out to compare these five measures in order to figure out whether they tap on fundamentally different aspects of children's vocal development, or whether they yield identical results, which would indicate that they are mere variations on the same theme.

In addition to the differences in how to compute these five measures, the size of the sample has not been an issue at stake in the literature on babbling onset: How many utterances or syllables are needed in order to reliably compute the measures? This question has been largely neglected in the past: researchers set an arbitrary number of syllables/utterances to be counted or an arbitrary number of minutes of recording was used as the yardstick. Although the reliability and validity of the sampling procedures were not tested, the results of different studies and measures all pointed in the same direction: for typically developing children the onset of babbling should occur certainly before the age of 0;11. The robust finding of a typical age range for babbling onset stimulated research on populations that are delayed in babbling onset compared to typically developing peers. An often replicated finding is that infants with profound hearing impairment show delays in the onset of babbling (Koopmans-van Beinum *et al.*, 2001; Oller & Eilers, 1988; Oller *et al.*, 1985; Stoel-Gammon & Otomo, 1986). Moreover, a delay in the onset of babbling in the absence of other diagnosed impairments was established to be a good predictor of later speech and language disorders (Oller *et al.*, 1999). Because findings such as these reveal age of babbling onset as a diagnostic marker and the calculation of it as a possible tool in the hands of clinicians and other health-care workers, care should be taken regarding the reliability of the sampling procedures with which babbling onset measures are applied. One wants to have confidence that the application of babbling onset measures to samples of vocalizations from different children yields a correct and reliable comparison between these children. Only if measures can be applied reliably can their results be of theoretical or clinical relevance.

Recently the issue of sampling and sample size was put on the agenda by Tomasello and Stahl (2004) and Rowland *et al.* (2008), though it was already quite pertinently a focus of attention in Richards' and Malvern's work on making the type/token ratio independent of sample size (Malvern *et al.*, 2004). Hence, before comparing the five measures for babbling onset, we investigated the questions: Can we establish a minimal sample size that permits a reliable identification of the age at onset of babbling with each of those measures? How big should that sample be?

Experiment 1 showed that computing a measure one single time on a relatively small sample is a hazardous undertaking: it was illustrated how selecting one sample of 50 utterances, and computing CBR^{utt} one single time, placed the onset of babbling as early as 0;8, or not yet at 1;0, for the same child. The aim of Experiment 1 was to find out if a sample size can be found that allows the computation of the five measures in just one single run with at least 95% confidence. Through the application of a bootstrapping procedure with increasing sample sizes, it became clear that the sheer amount of data available for the subjects can sometimes determine whether

the babbling onset border is crossed or not with sufficient confidence. As illustrated in Figure 1, if more data are available a particular subject can reach the critical threshold and another subject can fail to reach that threshold simply because there are more data collected of the former (often crucially depending on the volubility of the child at the time of the recording). Moreover, the bootstrapping exercise revealed that the number of items required for reliably computing the onset of babbling is relatively elevated: depending on the measure 300 to 500 items (syllables/utterances) are needed. Since especially at the younger ages that number of items is not collected easily in a reasonable amount of time, and since that amount of data is even harder to collect in a clinical setting, an alternative strategy was developed in Experiment 2.

The procedure proposed in Experiment 2 takes into account two important recommendations arrived at in the first experiment: (1) it is of crucial importance that if samples of relatively small size are used an iterative process of computing the measure for determining babbling onset is implemented; and (2) it is of critical importance that the same sample size is used for all subjects. The second bootstrapping experiment revealed that – dependent on the measure – 25 to 75 items suffice to reliably determine the onset of babbling. However, these small sample sizes require that the computation of the measure is repeated a sufficient number of times. This process is hardly feasible given a paper-and-pencil approach. But given present-day computational power, 1,000 iterations with random sampling over the entire dataset can be done in just a few seconds.

The outcomes of this study clearly indicate that the sample sizes used in previous research were sometimes too small to provide reliable estimates of the age at babbling onset for all children in a group, especially because babbling onset measures were only computed one single time per transcript. Nevertheless, as already indicated, reports are highly consistent in pinpointing the onset of babbling between six and eleven months of age. Those studies did not implement computationally costly strategies such as the ones advocated in this paper. How can this apparently startling contradiction be explained?

An explanation can perhaps be found in the opposition between a rigorous methodological framework and an intuitive assessment of children's vocal development. On the one hand, studies of prelexical vocal behaviour suffer from sparse data. The volubility of infants differs from occasion to occasion (Molemans *et al.*, 2010) and everyone who has ever made recordings of spontaneous vocal behaviour can testify that it sometimes takes a lot of time and patience to collect even a small sample of vocalizations. Experiment 1 shows that judging the passing of a babbling onset threshold on the basis of one single small sample is not without risks.

However, the results of Experiment 2 indicate that for all the children involved in this study, there is at least one observation session between 0;6 and 0;11 in which the measures reach a peak value that is so elevated that even small samples can (possibly) permit accurate determination of babbling onset. For instance, for all children in this study, CBR^{utt} reaches such a peak value in the period considered that has a median of 0·99 (range: 0·32 – 1·72), which indicates that almost every utterance contains a canonical syllable. Thus, with a longitudinal corpus containing (even relatively small) monthly samples and computing (one of) the measure(s) only once, the chances are that the age of babbling onset will be credited before 0;11, even though that age at babbling onset may not be fully accurate as with the procedure proposed in this paper.

What is reassuring in this respect is that in a study involving parents of very low economic status (Oller, Eilers & Basinger, 2001) found that '90% or more of the parents are aware, without any training at all, whether their infants are in the canonical stage of vocal development'. And they add that there probably is an 'intuitive awareness' in every parent of the important milestones of speech and language development of their children. This may suggest that in order to make a fully accurate assessment of the onset of babbling, quantitative means, such as the procedure proposed in this study, should be complemented with a parental questionnaire such as the one used by Oller *et al.* (2001). But it should be kept in mind that in the procedure of Oller and colleagues, parents only judge the presence of canonical syllables, and do not judge the surpassing of a particular quantitative threshold, while the procedure proposed in this paper computes a quantitative measure – Does the number of canonical syllables surpass a particular threshold? – and not just the presence of canonical syllables.

REFERENCES

Baayen, H. (2008). *Analyzing linguistic data*. Cambridge: Cambridge University Press.

Chapman, K., Hardin-Jones, M., Schulte, J. & Halter, K. (2001). Vocal development of 9-month-old babies with cleft palate. *Journal of Speech, Language, and Hearing Research* **44**, 1268–83.

Clements, G. (1990). The role of the sonority cycle in core syllabification. In J. Kingston & M. Beckman (eds), *Papers in laboratory phonology I: Between the grammar and physics of speech*, 283–333. Cambridge: Cambridge University Press.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* **7**, 1–26.

Fagan, M. K. (2009). Mean Length of Utterance before words and grammar: Longitudinal trends and developmental implications of infant vocalizations. *Journal of Child Language* **36**, 495–527.

Heilman, J., Nockerts, A. & Miller, J. (2010). Language sampling: Does the length of the transcript matter? *Language, Speech and Hearing Sciences in Schools* **41**, 393–404.

Hutchins, T., Brannick, M., Bryant, J. & Silliman, R. (2005). Methods for controlling amount of talk: Difficulties, considerations and recommendations. *First Language* **25**, 347–63.

Klee, T. & Fitzgerald, M. (1985). The relation between grammatical development and mean length of utterance in morphemes. *Journal of Child Language* **12**, 251–69.
Koopmans-van Beinum, F., Clement, C. & van den Dikkenberg-Pot, I. (2001). Babbling and the lack of auditory speech perception: A matter of coordination? *Developmental Science* **4**, 61–70.
Koopmans-van Beinum, F. J. & van der Stelt, J. (1986). Early stages in the development of speech movements. In B. Lindblom & R. Zetterstrom (eds), *Precursors of early speech*, 37–50. New York: Stockton.
Landis, J. & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–74.
Lynch, M., Oller, D., Steffens, M., Levine, S., Basinger, D. & Umbel, V. (1995). Onset of speech-like vocalizations in infants with Down syndrome. *American Journal of Mental Retardation* **100**, 68–86.
MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum.
Malvern, D., Richards, B., Chipere, N. & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Houndmills: Palgrave Macmillan.
Molemans, I., Van Severen, L., van den Berg, R., Govaerts, P. & Gillis, S. (2010). Spraakzaamheid van Nederlandstalige baby's en peuters: Longitudinale spontane spraakdata. *Logopedie* **23**, 12–23.
Morris, S. (2010). Clinical application of the Mean Babbling Level and Syllable Structure Level. *Language, Speech and Hearing Sciences in Schools* **41**, 223–30.
Nathani, S., Ertmer, D. J. & Stark, R. E. (2006). Assessing vocal development in infants and toddlers. *Clinical Linguistics & Phonetics* **20**(5), 351–69.
Oller, D. K. (1980). The emergence of the sounds of speech in infancy. In G. H. Yeni-Komshian, J. F. Kavanagh & C. A. Ferguson (eds), *Child phonology. Volume 1: production*, 93–112. New York: Academic Press.
Oller, D. K. (2000). *The emergence of the speech capacity*. Mahwah, NJ: Lawrence Erlbaum.
Oller, D. K. & Eilers, R. (1988). The role of audition in infant babbling. *Child Development* **59**, 441–49.
Oller, D. K., Eilers, R. & Basinger, D. (2001). Intuitive identification of infant vocal sounds by parents. *Developmental Science* **4**, 49–60.
Oller, D. K., Eilers, R., Bull, D. & Carney, A. (1985). Pre-speech vocalizations of a deaf infant: A comparison with normal metaphonological development. *Journal of Speech and Hearing Research* **28**, 47–63.
Oller, D. K., Eilers, R., Neal, A. & Schwartz, H. (1999). Precursors to speech in infancy: The prediction of speech and language disorders. *Journal of Communication Disorders* **32**, 223–45.
Oller, D. K., Eilers, R., Steffens, M., Lynch, M. & Urbano, R. (1994). Speech-like vocalizations in infancy: An evaluation of potential risk factors. *Journal of Child Language* **21**, 33–58.
Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of Child Language* **14**, 201–209.
Roug, L., Landberg, I. & Lundberg, L.-J. (1989). Phonetic development in early infancy: A study of four Swedish children during the first eighteen months of life. *Journal of Child Language* **16**, 19–40.
Rowland, C., Fletcher, S. & Freudenthal, D. (2008). How big is big enough? Assessing the reliability of data from naturalistic samples. In H. Behrens (ed.), *Corpora in language acquisition research: History, methods, perspectives*, 1–24. Amsterdam: Benjamins.
Rvachew, S., Creighton, D., Feldman, N. & Sauve, R. (2001). Acoustic-phonetic description of infant speech samples: Coding reliability and related methodological issues. *Acoustics Research Letters Online* **3**, 24–28.
Schauwers, K., Gillis, S., Daemers, K., De Beukelaer, C. & Govaerts, P. (2004). The onset of babbling and the audiological outcome in cochlear implantation between 5 and 20 months of age. *Otology and Neurotology* **25**, 263–70.

Stark, R. E. (1980). Stages of speech development in the first year of life. In G. H. Yeni-Komshian, J. F. Kavanagh & C. A. Ferguson (eds), *Child phonology. Volume 1: production*, 163–73. New York: Academic Press.
Stoel-Gammon, C. (1989). Prespeech and early speech development of two late talkers. *First Language* **9**, 207–224.
Stoel-Gammon, C. & Otomo, K. (1986). Babbling development of hearing impaired and normally hearing subjects. *Journal of Speech and Hearing Disorders* **51**, 33–41.
Tomasello, M. & Stahl, D. (2004). Sampling children's spontaneous speech: How much is enough? *Journal of Child Language* **31**, 101–121.
van der Stelt, J. & Koopmans-van Beinum, F. (1986). The onset of babbling related to gross motor development. In B. Lindblom & R. Zetterström (eds), *Precursors of early speech*, 163–73. New York: Stockton Press.
Woods, A., Fletcher, P. & Hughes, A. (1986). *Statistics in language studies*. Cambridge: Cambridge University Press.
Zink, I. & Lejaegere, M. (2002). *N-CDIs Lijsten voor Communicatieve Ontwikkeling*. Leuven: ACCO.